Investigating the effects of race on vote choice with frequentist and machine learning logistic regression models

Talia Fabregas 1008350213

June 2025

Abstract

This assignment uses logistic regression models to investigate the impact of race and its interactions with demographic and geographic variables on vote choice in the 2024 U.S. Presidential Election. After applying both frequentist and machine learning (ML) logistic regression models to the 2024 Congressional Election Study (CES) dataset, I found that race and its interaction with region significantly improve the fit of frequentist logistic vote choice models and an ML model that included race and its interactions with gender, education, region, and the urban-rural divide achieved 91% test accuracy. Although unconventional, removing the race main effect but keeping its interaction terms did not worsen test accuracy; this suggests that the standalone effect of race on voting behavior might be declining but its interactions with demographic and geographic factors remain important. ML models showed higher precision, recall, and F1-score for 2024 CES respondents that voted for Harris compared to Trump due to the under-representation of Trump voters in the sample. These results are consistent with existing research that suggests race influences voting behavior in a complex, demographic and geographic-dependent way and incorporating interaction effects can improve modeling accuracy.

Code and data can be found using this link:
https://colab.research.google.com/drive/1pXp_rB2Y8yM7iQAJ1rQmbBPcmb3TIOA9?usp=sharing

1 Introduction

In 2024, U.S. President Donald Trump's voter coalition was more racially and ethnically diverse than in 2020 or 2016 [1]. This assignment (turned mini, informal paper) was inspired by a tweet by Canadian political commentator J.J. McCullough that uses the term "racedep" to describe Americans voting in less racially predictable ways and the publication titled "The Geography of Racially Polarized Voting: Calibrating Surveys at the District Level" by Yale University Researcher Shiro Kuriwaki and colleagues [2]. As of now, racedep is not really an academic theory; I scoured google scholar and the UofT library search for research articles about "race depolarization" or "racedep" published between the November 2024 U.S. presidential election and now. I had little success, so for the context of this assignment, I will stick with McCullough's definition of racedep.

In their 2023 study, which involved conducting large-scale sample surveys with aggregate demographic and election data, Kuriwaki et al. found that national-level differences across racial groups explain 60% of differences in voting behavior at the district level, while geography explains 30% [2]. A Pew Research Center study of the 2024 election based on a nationally representative sample of validated voters found that Trump made double-digit gains with Hispanic and Asian voters and increased his support among Black voters by 8 points; older white voters without a college education were also more likely to support Trump [1]. Race still explains vote choice more than geography, but its role as a deciding factor in vote choice might be declining relative to other factors including the educational divide and the rural-urban gap [2]. Since his first Presidential campaign in 2016, Trump has made significant gains with voters of color, but also widened the educational divide and the rural-urban gap [1]. Congressional districts in the South and Midwest tend to have the highest levels of racially-polarized voting; this suggests that the effects of region may vary across different racial groups [2].

I will start with a frequentist approach, inspired by Kuriwaki et al. [2]. I will fit a more complex logistic regression model that includes race as a main effect and an interaction between race and region and measure the significance of race on its own in the 2024 CES datasets. I will then use likelihood ratio tests to determine how much including race and the race-region interaction improves model fit. I will also try out a machine learning (ML). I will also use a Machine Learning (ML) approach; this will go further than my frequentist approach and examine whether adding in the interactions between race and gender and race and education can improve model performance (test accuracy). My ML approach will involve training more complex logistic models to classify Trump and non-Trump (Harris) voters in the 2024 CES dataset, then reporting each model's performance. My mid-term paper will build on this, incorporate and review more existing research, and present more in-depth modeling and results.

2 Data

I used the 2024 Congressional Election Study (CES) Common Content Dataset, obtained from Harvard Dataverse. The CES is 60,000 respondent nationally representative survey that aims to measure voter registration and voting behavior leading up to and immediately after U.S. presidential and midterm elections [3]. The dataset was cleaned to include a subset of columns (CC24_364a, birthyr, gender4, race, educ, region, urbancity, religpew, pid3) and to remove all rows with NULL values for a more streamlined modeling process using the `scikit-learn` package. I subsetted it to only include respondents who voted for either Trump or Harris in 2024; logistic regression is used to model binary outcomes, so third-party voters were not included. I created the `age_bracket` variable using `birthyr` and the `vote_trump` variable by using the `CC24_364a` variable (who the respondent voted for in 2024) and wrote code to make it equal to 1 if the respondent voted for Trump and 0 otherwise.

3 Methods

My outcome variable is support for Trump; the binary outcome variable vote_trump is equal to one of the CES respondent voted for President Donald Trump and 0 if the respondent voted for former Vice President Kamala Harris in the 2024 Presidential Election. I will use both frequentist and machine learning logistic regression models to measure it.

4 Models

I explored two modeling approaches: frequentist and machine learning (ML). I took STA303 (Methods of Data Analysis II), CSC311 (Introduction to Machine Learning), and STA414 (Probabilistic Machine Learning) so I saw this assignment as an opportunity to apply what I learned in those two classes to vote choice modeling. Given the binary nature of US elections (two major party nominees, third-party candidates receive a small percentage of the vote and are extremely unlikely to win any electoral college votes), logistic regression can be used to classify CES survey respondents as Trump or non-Trump (Biden in 2020, Harris in 2024) voters. The purpose of this assignment was to try out different logistic regression models and approaches to measure voting behavior.

4.1 Frequentist logistic regression models

I used the `statsmodels.formula.api` Python package to fit each of the logistic models outlined below on 2024 CES survey dataset.

Model 0 (Full model):

$$\begin{aligned} p(\text{vote\_trump} = 1) = \sigma\big(&\beta_0 + \beta_1 \cdot \text{race} + \beta_2 \cdot \text{gender} + \beta_3 \cdot \text{educ} + \beta_4 \cdot \text{age\_bracket} + \beta_5 \cdot \text{region} \\ &+ \beta_6 \cdot \text{urbancity} + \beta_7 \cdot \text{religion} + \beta_8 \cdot \text{party\_id} + \beta_9 \cdot (\text{race} \times \text{region})\big) \end{aligned}$$

Model 1 (omit the race × region interaction):

$$\begin{aligned} p(\text{vote\_trump} = 1) = \sigma\big(&\beta_0 + \beta_1 \cdot \text{race} + \beta_2 \cdot \text{gender} + \beta_3 \cdot \text{educ} + \beta_4 \cdot \text{age\_bracket} + \beta_5 \cdot \text{region} \\ &+ \beta_6 \cdot \text{urbancity} + \beta_7 \cdot \text{religion} + \beta_8 \cdot \text{party\_id}\big) \end{aligned}$$

Model 2 (omit all race-related predictors):

$$\begin{aligned} p(\text{vote\_trump} = 1) = \sigma\big(&\beta_0 + \beta_1 \cdot \text{gender} + \beta_2 \cdot \text{educ} + \beta_3 \cdot \text{age\_bracket} + \beta_4 \cdot \text{region} \\ &+ \beta_5 \cdot \text{urbancity} + \beta_6 \cdot \text{religion} + \beta_7 \cdot \text{party\_id}\big) \end{aligned}$$

4.2 Machine learning logistic regression models

I used the `scikit-learn` Python package to split the 2020 and 2024 CES data into training and test datasets and fit and evaluate the performance of each of the models outlined below. In past projects (STA302 and STA304), I found that the effects of race may also vary based on gender and highest level of education completed. Including the race × gender and race × education interactions in my `statsmodels.formula.api` models was impractical because there are too many sparse combinations. However, I was able to use `scikit-learn` to split the CES data into training and test sets, train various logistic regression models to identify Trump voters, and report test accuracy.

Model A:

$$\begin{aligned} p(\text{vote\_trump} = 1) = \sigma\Big(&\beta_0 + \beta_1 \cdot \text{race} + \beta_2 \cdot \text{gender} + \beta_3 \cdot \text{educ} + \beta_4 \cdot \text{age\_bracket} + \beta_5 \cdot \text{region} \\ &+ \beta_6 \cdot \text{urbancity} + \beta_7 \cdot \text{religion} + \beta_8 \cdot \text{party\_id} + \beta_9 \cdot (\text{race} \times \text{gender}) \\ &+ \beta_{10} \cdot (\text{race} \times \text{educ}) + \beta_{11} \cdot (\text{race} \times \text{region}) + \beta_{12} \cdot (\text{race} \times \text{urbancity})\Big) \end{aligned}$$

Model B (omit race main effect):

$$\begin{aligned} p(\text{vote\_trump} = 1) = \sigma\Big(&\beta_0 + \beta_1 \cdot \text{gender} + \beta_2 \cdot \text{educ} + \beta_3 \cdot \text{age\_bracket} + \beta_4 \cdot \text{region} \\ &+ \beta_5 \cdot \text{urbancity} + \beta_6 \cdot \text{religion} + \beta_7 \cdot \text{party\_id} + \beta_8 \cdot (\text{race} \times \text{gender}) \\ &+ \beta_9 \cdot (\text{race} \times \text{educ}) + \beta_{10} \cdot (\text{race} \times \text{region}) + \beta_{11} \cdot (\text{race} \times \text{urbancity})\Big) \end{aligned}$$

Model C (omit regional interactions):

$$\begin{aligned} p(\text{vote\_trump} = 1) = \sigma\Big(&\beta_0 + \beta_1 \cdot \text{race} + \beta_2 \cdot \text{gender} + \beta_3 \cdot \text{educ} + \beta_4 \cdot \text{age\_bracket} + \beta_5 \cdot \text{region} \\ &+ \beta_6 \cdot \text{urbancity} + \beta_7 \cdot \text{religion} + \beta_8 \cdot \text{party\_id} + \beta_9 \cdot (\text{race} \times \text{gender}) \\ &+ \beta_{10} \cdot (\text{race} \times \text{educ}) + \beta_{11} \cdot (\text{race} \times \text{region}) + \beta_{12} \cdot (\text{race} \times \text{urbancity})\Big) \end{aligned}$$

Model D (omit the race × gender and race × education interactions):

$$\begin{aligned} p(\text{vote\_trump} = 1) = \sigma\Big(&\beta_0 + \beta_1 \cdot \text{race} + \beta_2 \cdot \text{gender} + \beta_3 \cdot \text{educ} + \beta_4 \cdot \text{age\_bracket} + \beta_5 \cdot \text{region} \\ &+ \beta_6 \cdot \text{urbancity} + \beta_7 \cdot \text{religion} + \beta_8 \cdot \text{party\_id} + \beta_9 \cdot (\text{race} \times \text{educ}) \\ &+ \beta_{10} \cdot (\text{race} \times \text{region}) + \beta_{11} \cdot (\text{race} \times \text{urbancity})\Big) \end{aligned}$$

5 Results

5.1 Frequentist

Likelihood ratio tests were used to compare the reduced and full logistic regression model fits on the 2024 CES data. As shown in Figure 1, the first likelihood ratio test (comparing Model 0 to Model 1) resulted in a likelihood ratio (LR) statistic of 47.14, degrees of freedom difference of 21, and p-value of approximately $0.0009 << 0.05$. This provides strong evidence against the null hypothesis that the simpler model (Model 1) is sufficient and fits the 2024 CES data as well as the more complex model (Model 0). The additional 21 race × gender dummy predictors included in model 0 explain significant variation in 2024 vote choice that is not captured by Model 1. These results support retaining Model 0.

2024

```
[ ] # model 0 vs model 1
    print(likelihood_ratio(ces24_model0, ces24_model1))
```

```
LR stat: 47.1416, full df - reduced df: 21, p-value: 0.0008987791605244988
```

```
[ ] # model 1 vs model 2
    print(likelihood_ratio(ces24_model1, ces24_model2))
```

```
LR stat: 38.3388, full df - reduced df: 7, p-value: 2.6123141803058303e-06
```

```
[ ] # model 0 vs model 2
    print(likelihood_ratio(ces24_model0, ces24_model2))
```

```
LR stat: 85.4804, full df - reduced df: 28, p-value: 9.91846926812201e-08
```

Figure 1: Likelihood ratio test results for the frequentist logistic models on the 2024 CES data. The first likelihood ratio test (model 0 vs model 1) has a p-value of $0.0009 < 0.05$; this indicates strong evidence against the null hypothesis at the $\alpha = 0.05$ significance level and that the race $\times$ region interaction should not be removed from the model. The second likelihood ratio test (model 1 vs model 2) has a p-value of $2.16e^{-6} << 0.5$; the model that includes race but not the race $\times$ region interaction still fits the data better than the model that drops the race main effect. The third likelihood ratio test (which is redundant at this point) has a p-value of approximately $9.92e^{-8} << 0.5$; removing the 28 additional race and race $\times$ gender parameters in the full model worsens model fit.

5.2 Machine learning

Each ML logistic model outlined in section 4.2 was trained and tested on t 2024 CES dataset using the `train_test_split` function in the `scikit-learn` package. The default `scikit-learn` split size was used; 75% of the dataset was used for training and 25% was used for testing [4]. I fit each model using the `LogisticRegression` function of the `scikit-learn` package. In this section, I am building on my results from section 5.1 which show that a logistic model that includes the race main effect and the race $\times$ region interaction fits the 2024 CES dataset better than the model that omits these predictors. This section builds on the 2024 Pew Research Study that suggests interactions between the effects of race and gender and race and education on voting behavior [1]. For each model outlined in section 4.2, I will report the overall test accuracy and precision, recall, and F1-score for both the negative class (respondents who voted for Harris) and the positive class (respondents who voted for Trump) using the `scikit-learn classification_report` function. Overall test accuracy measures how often the model correctly predicts vote choice (identifies Trump and Harris voters) across the entire test dataset, precision is the proportion of respondents predicted to belong to a given class (Trump or Harris voters) that actually do, recall is the proportion of respondents who belong to a given class (Trump or Harris voters) that the model correctly identified as such, and F1-score is the weighted mean of precision and recall [4].

As shown in Table 1, Model A had a 91% overall test accuracy on the 2024 CES survey dataset; this means that it correctly classified 91% of respondents as Trump or Harris voters in the test set. Table 2 shows that removing the race main effect and keeping all of the race-related interaction simplified the model (removed 7 parameters) without compromising test accuracy. Kuriwaki et al. suggested that while race is still a predictor of vote choice, its strength relative to other factors such as education and geography might be declining [2]. For this reason, I wanted to examine how a logistic model that drops the race main effect, but keeps its interactions with education, gender, region, and whether the respondent lives in an urban or rural area performs. After completing this part of the assignment, I learned that this is not a standard statistical practice but I am still including it because it shows a way in which the model *can* be simplified without reducing its ability to classify Trump and Harris voters.

All five models outlined in section 4.2 had lower precision, recall, and F1-scores for the positive class (Trump voters) than for the negative class (Harris voters); this means that the models were better at identifying Harris voters than Trump voters. The support for the positive class is also lower than the support for the positive class; this is unsurprising because Trump voters are under-represented in the 2024 CES survey dataset [3]. Model A has the smallest gaps in precision, recall, and F1-score between class 0 (Harris voters) and class 1 (Trump voters at 0.04, 0.14, and 0.09 respectively. Model C and Model D show the most significant class 0 recall and F1-score drops compared to Model A. Models C and D also have the largest precision, recall, and F1-score gaps between class 0 and class 1.

Class	Precision	Recall	F1-score	Support
0 (Harris voters)	0.92	0.96	0.94	1587
1 (Trump voters)	0.88	0.82	0.85	725
Overall Test Accuracy	0.91 (Total support: 2312)			
Macro avg	0.90	0.89	0.89	2312
Weighted avg	0.91	0.91	0.91	2312

Table 1: This table shows the **classification performance** of the full machine learning logistic regression model (Machine Learning **Model A**) in predicting `vote_trump` on the 2024 CES test data. Class 0 refers to the negative class (voted for Harris) and class 1 is refers to the positive class (voted for Trump). Precision, recall, F1-scores, and class support are reported. Notably, the model had higher precision, recall, and F1-scores in the negative class (0.92, 0.96, and 0.94 vs 0.88, 0.82, and 0.85). This means that the model is better at accurately identifying Harris voters than Trump voters. Model A has the smallest precision, recall, and F1-score differences between the positive and negative class. As shown near the bottom of the table, Model A has a 91% overall test accuracy

Class	Precision	Recall	F1-score	Support
0 (Harris voters)	0.91	0.96	0.93	1582
1 (Trump voters)	0.89	0.80	0.84	730
Overall Test Accuracy	0.91 (Total support: 2312)			
Macro avg	0.90	0.89	0.89	2312
Weighted avg	0.90	0.91	0.90	2312

Table 2: This table shows the **classification performance** of Model B. Class 0 refers to the negative class (voted for Harris) and class 1 is refers to the positive class (voted for Trump). Precision, recall, F1-scores, and class support are reported. Overall, model B has the same overall test accuracy as model A (91%) and nearly identical precision and recall. This suggests that removing the race main effect while keeping the race × gender, race × education, race × region and race × urban interactions can reduce model complexity without compromising performance. I have since learned that removing a main effect while keeping its interactions is not standard practice in statistics; I did it for the sole purpose of testing race depolarization and building a vote choice model that balances performance with complexity.

Class	Precision	Recall	F1-score	Support
0 (Harris voters)	0.90	0.94	0.92	1600
1 (Trump voters)	0.88	0.77	0.82	712
Overall Test Accuracy	0.89 (Total support: 2312)			
Macro avg	0.89	0.86	0.87	2312
Weighted avg	0.90	0.90	0.89	2312

Table 3: This table shows the **classification performance** of Model C. Class 0 refers to the negative class (voted for Harris) and class 1 refers to the positive class (voted for Trump). Precision, recall, F1-scores, and class support are reported. This model performs better on class 0 than on class 1. Overall, Model C saw a two percentage point decrease in overall test accuracy compared to Model A and Model B (80% vs 91%) and had lower precision, recall, and F1-scores across both classes. These results are consistent with section 5.1; the logistic model that includes the race × region interaction fits the data better than the one that omits it. Additionally, Model C has a larger gap in precision, recall, and F1-scores between the positive and negative classes compared to Model A.

Class	Precision	Recall	F1-score	Support
0 (Harris voters)	0.90	0.95	0.93	1604
1 (Trump voters)	0.88	0.76	0.82	708
Overall Test Accuracy	0.89 (Total support: 2312)			
Macro avg	0.89	0.86	0.87	2312
Weighted avg	0.89	0.89	0.89	2312

Table 4: This table shows the **classification performance** of Model D. Class 0 refers to the negative class and class 1 refers to the positive class. Precision, recall, and F1-score for each class are reported. Model D performs better on class 0 than class 1 (0.90, 0.95, and 0.93 vs 0.88, 0.76, and 0.82 precision, recall, and F1-score respectively). Model D has the largest difference (0.19) between class 0 and class 1 recall (0.95 vs 0.76) and an 89% overall test accuracy. While Model D's overall test accuracy is only two percentage points lower than that of Model A, it shows a drop in performance on the positive class and larger performance gaps between the positive and negative classes.

6 Discussion

6.1 Consistency with existing research and relation to the race depolarization theory

My frequentist results in Section 5.1 show that a logistic regression that includes the interaction between race and region fits the 2024 CES dataset better than a model that excludes it. A model that excludes the race main effect does not fit the data as well as a model that includes it. This is consistent with the findings of Kuriwaki et al. [2]; race is still a significant predictor of voting behavior, but the effects of race can vary based on geographic factors. In Section 5.2, I found that removing race as a main effect but keeping the race × education, race × gender, race × region and race × urban-city interaction effects reduced model complexity by removing 7 race dummy variables without compromising performance. Precision, recall, and F1 score for both classes (Harris and Trump voters) did not change significantly; the largest change was a two percentage point decrease in Trump voter recall from 82% to 80%. However, the overall test accuracy remained the same at 91%. This suggests that in the 2024 presidential election, race might have been a weaker standalone predictor of vote choice than in the past, but its effects vary by gender, highest level of educational attainment, and geography. My finding that Model A has the smallest precision, recall, and F1-score gap between class 0 and class 1 and that Models C and D have the largest performance drop on the positive class in Section 5.2 is consistent with Section 5.1, Kuriwaki et al., and the Pew research study of the 2024 election. The model that includes the race × region interaction fits the data better than the model that omits it and including the race × gender, race × education, race × region, and race × urbancity interactions improved the model's accuracy, especially its ability to correctly identify Trump voters.

6.2 Weaknesses and limitations

One weakness of my modeling assignment is that I only fit models on the 2024 CES dataset. It is unclear how well my model would generalize, perform on unseen data, and predict future voting behavior in the United States. Another weakness of my results is that my models were trained on a dataset that under-sampled Trump voters; the vast majority of 2024 CES respondents voted for Harris, despite the fact that Trump won both the popular vote and the electoral college in the 2024 presidential election.

Another weakness that this assignment has is its lack of comparison with previous presidential elections. Although my ML results in Section 5.2 show that removing the race main effect while keeping the race-related interaction effects can reduce model complexity without compromising its performance on the 2024 CES survey dataset, I have not yet investigated if or how this applies to past elections. Additionally, my frequentist likelihood ratio results show that removing the race main effect and the interaction effect between race and region worsens model fit on the 2024 CES survey dataset, but I have not compared this with the other elections that Trump ran in (2016 and 2020) or the pre-Trump era. I can compare the significance of race as a stand-alone predictor and its interactions with various demographic and geographic variables in previous elections to further investigate whether the effects of race have been declining relative to other variables or the extent to which its effect is contextual, context-specific, or in conjunction with other factors.

6.3 Next steps

My goal is to extend the work I did for this assignment into a paper. In my mid-term paper, I will consider including data from the 2016 US Presidential election (Trump's first election) and the pre-Trump era (2008 and 2012). I will also consider incorporating the findings of Algara et al. into a logistic vote choice model [5]. Algara et al. examined the correlation between incumbent presidential approval ratings and popular vote share and correctly predicted that Trump would win the 2024 election [5]. They performed statistical analyses and modeling, and found that an individual's approval (or disapproval) of the incumbent president is highly predictive of their likelihood of voting for the president's party's nominee [5]. This assignment primarily focused on the findings of Kuriwaki et al. and how race predicts vote choice, but my midterm paper will explore how incorporating additional predictors and increasing model complexity can improve performance.

My goals for my midterm paper and second modeling assignment include:

- Including visualizations to illustrate my exploratory data analysis process and how predictors were chosen.
- Incorporating the work of Algara et al. to build a vote choice model with improved accuracy on Trump voters in various CES survey datasets.
- Testing the significance predictors related to approval of the incumbent president and his party's brand.
- Exploring different ways to improve model performance on Trump voters and how to address survey non-response/selection bias.
- Using the 2008-2024 Presidential Election CES survey datasets to investigate how the role of race as a predictor of vote choice has changed over time.
- Explore how to use regularization and gradient descent.

References

[1] H. Hartig, S. Keeter, A. Daniller, and T. Van Green, "Behind Trump's 2024 Victory, a More Racially and Ethnically Diverse Voter Coalition," Pew Research Center, Jun. 26, 2025. [Online]. Available: https://www.pewresearch.org/politics/2025/06/26/behind-trumps-2024-victory-a-more-racially-and-ethnically-diverse-voter-coalition

[2] S. Kuriwaki, J. Grumbach, A. Sadin, and E. Schickler, "The Geography of Racially Polarized Voting: Calibrating Surveys at the District Level," Institution for Social and Policy Studies, Yale University, ISPS22-28, 2023. [Online]. Available: https://isps.yale.edu/research/publications/23-28

[3] B. Schaffner, M. Shih, S. Ansolabehere, and J. Pope, "Cooperative Election Study Common Content, 2024." Harvard Dataverse, 2025. doi: doi:10.7910/DVN/X11EP6.

[4] F. Pedregosa et al., "3.4 Metrics and Scoring: quantifying the quality of predictions," scikit-learn: Machine Learning in Python, scikit-learn developers, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/model_evaluation.html

[5] C. Algara and K. Hale, "Does Racial Attitude Influence the 2020 Vote for Trump? Evidence from a National Probability Sample of White Voters," Political Behavior, vol. 45, no. 1, pp. 193–216, Mar. 2023. [Online]. Available: https://doi.org/10.1007/s11109-021-09727-5